Corpus-based part-of-speech disambiguation of Persian

Mosavi Miangah Tayebeh
English Language Department
Payame Noor University, Yazd, Iran
Email: mosavit@pnu.ac.ir; mosavit@hotmail.com

***Abstract*—In this paper we introduce a method for part-of-speech disambiguation of Persian texts, which uses word class probabilities in a relatively small training corpus in order to automatically tag unrestricted Persian texts. The experiment has been carried out in two levels as unigram and bi-gram genotypes disambiguation. Comparing the results gained from the two levels, we show that using immediate right context to which a given word belongs can increase the accuracy rate of the system to a high degree.**

***Index Terms*—genotype, machine translation, part of speech disambiguation, word class probabilities**

I.INTRODUCTION

In linguistics, the term 'corpus' refers to a relatively large number of raw or annotated words in the body of text. Computational linguists recently turned into corpus-based approaches for solving various linguistic problems such as phrase recognition [6], word sense disambiguation [14], building dictionaries, morphological analysis and automatic lemmatization [10] and [13], language teaching [5], machine translation [19], information retrieval [2] and some other problems. Naturally, preparing a tagged or annotated corpus from different points of view has been of great significance for anyone who involves in computational linguistics career. Constructing such an annotated corpus has already been done for many languages including English, Czech, German, Hungarian, French and Arabic to name a few. Automatic part-of-speech (below POS) disambiguation of a large corpus has been studied applying different approaches that we will go through some of them in what follows.

To start with Persian, it should be said that corpus-based approaches for text analysis have a rather short history in Persian language. The only serious attempt ever taken in this connection is constructing an interactive POS tagging system developed by [1]. In their project they followed the methods proposed in [18]. It is based on the hypothesis that syntactic behavior is reflected in co-occurrence patterns. Therefore, the similarity between two words will be measured with respect to their syntactic behaviors to their left side by the degree to which they share the same neighbors on the left. So, the word types are recognized according to their distributional similarity (their similarity in terms of sharing the same neighbors), and then each category can be manually tagged [1]. In this way a grammatically tagged corpus of Persian was created making up of 45 tags which have designed with reference to the categories normally introduced in dictionaries. Each tag is made up of one to five characters. In general, the accuracy of this kind of distributional POS tagging system proved to be 57.5%.

Brill presents a simple rule-based POS tagger, which automatically acquires its rules and tags with accuracy comparable to stochastic taggers [4]. Petasis, et al. study the performance of Transformation-Based Error Driven (TDED) learning for solving POS ambiguity in the Greek language, and examine its dependence on the thematic domain. For their work they trained the Brill tagger, [3], over relatively small-sized annotated Greek corpus and found its performance to be around 95% [17]. Daelemans, et al. introduce a memory-based approach to POS tagging. The POS tag of a word in a particular context is extrapolated from the most similar cases held in memory. Using this method, they obtain a tagging accuracy that is on a par with that of known statistical approaches, and with attractive space and time complexity properties when using IGTree, a tree-based formalism for indexing and searching huge case bases [7]. Kempe presents a method of constructing and applying a cascade consisting of a left and right-sequential finite-state transducer, T_1 and T_2, for POS disambiguation. In the process of POS tagging, every word is first assigned a unique ambiguity class that represents the set of alternative tags that this word can occur with. The sequence of the ambiguity classes of all words of one sequence is then mapped by T_1 to a sequence of reduced ambiguity classes where some of the less likely tags are removed. That sequence is finally mapped by T_2 to a sequence of single tags. [9]. Tzoukermann and Radev use word class for POS disambiguation. They investigate a direction for coming up with different kinds of probabilities based on paradigms of tags for given word. Their estimations are based not on the words, but on the sets of tags associated with a word. They claim that this approach gives a more efficient representation of the data in order to distinguish word POS. The accuracy of this method reached about 95% for POS disambiguation of unrestricted French text [20].

In this paper we present a method for POS disambiguation of Persian texts adopted the last-mentioned method above with some manipulations in order to conform with properties of Persian. In most statistically-based systems determining correct tags for words in the body of context the tendency is to count the frequencies corresponding to every word in a large annotated corpus and extract some probabilities relating to individual words rather than to tags or set of tags to which the individual words belong. Working with this kinds of probabilities may seem a desirable way to disambiguate such

DOI: 01.IJIT.01.01.5

a popular and widespread language as English for which publicly available annotated corpora like LOB corpus or Brown one exist. However, for a low density language as Persian for which resources in electronically readable form such as large annotated corpus are scarce, using words probabilities for POS disambiguation does not result satisfactorily. Hence, it seems more reasonable to use word class probabilities in a relatively small training corpus of Persian to be able to automatically tag a test corpus. We refer to this word class as 'genotype', series of tags based on morphological features (as it is used by [20]. A class of words may belong to the same genotype according to their possible tags. Then, when a word was disambiguated by some means, only one tag would be suitable for the context to which the word belongs and we refer to this correct tag as 'genotype decision'. In section 2 the training corpus used for POS disambiguation by word class as well as the tagset compiled for this purpose is introduced, and the experiment itself is discussed in section 3. Discussion of the results gained from the two levels of analysis (unigram and bigram levels) is presented in sections 4.

II.PERSIAN BACKGROUND

Persian is a member of synthetic language family. It means that in Persian a new word is to be created by adding prefix, suffix, infix or another noun, adjective, preposition or verb to the beginning or the end of the word or verb stem. In these cases the basic form of the word or verb stem usually is not broken [16]. Grammatical word order of Persian is SOV, although a relatively free word order is also possible, but not grammatically acceptable. In Persian every verb has two stems, present stem and past stem, and different inflectional forms of a verb is constructed either using the present stem, or the past one. This property of Persian verbs can be taken as an advantage in using stem dictionary method in automatic morphological analysis over the word-form dictionary or the other methods [14]. Consider the Persian infinitive *gftn* (to say) which has two stems as follows: *gui* (present stem), and *gft* (past stem) by which 38 different forms of the verb in different tenses can be constructed as follows:

gui , bgu , (b)guim , (b)guii , (b)guid , (b)guim , (b)guiid , (b)guind , gftm , gfti , gft , gftim , gftid , gftnd , gfth am , gfth ai , gfth aid , gfth aim , gfthaid , gfth and , gfth budm , gfth budi , gfth bud , gfth budim , gfth budid , gfth budnd , gfth ba:shm , gfth ba:shi , gfth ba:shd , gfth ba:shim , gfth ba:shid , gfth ba:shnd, xva:hm gft, xva:hi gft, xva:hd gft, xva:hin gft, xva:hid gft, xva:hnd gft [8].

In addition to verbs, many nouns, adjectives and adverbs in Persian are constructed from the present or past stem of the verbs. In these cases we name such words as "derivative" words as opposed to "concrete" (primary) words in which no verb stem involve. Nouns *da:nsh* (science) and *da:nshga:h* (university), adjectives *da:nshmnd* (scientist) and *bida:nsh* (ignorant) as well as the adverb *da:nshmnda:nh* (scholarly) have been constructed from present stem *da:n* which means "know" in English. Transliteration of Persian words used in this paper has been given in Appendix I).

Persian morphology is an affixal system consisting mainly of suffixes and a few prefixes. The nominal paradigm consists of a relatively small number of affixes. The verbal inflectional system is quite regular and can be obtained by the combination of prefixes, stems, inflections and auxiliaries [11]. An interesting point, here, is that the auxiliaries and modals in Persian are also subject to conjugation. Consider, for instance the two Persian verbs consisting of modal and auxiliary:

i)Mitva:nstm[1] brvm → [modal} + [verb] (I could go.)
[mi + tva:nst + m] + [b + rv + m]
[prefix + present stem + suffix]

ii) nrfth budid → [verb] + [auxiliary] (you had not gone.)
[n + rft + h] + [bud + id]
[prefix + past stem + suffix] + [stem + suffix]

in which both the auxiliary and modal take the affixes.

The elements within a noun phrase are linked by the enclitic particle called **ezafe**. This morpheme is usually an unwritten vowel, but it could also have an orthographic realization in certain phonological environments [11]. For example, when the last letter of the first noun is 'a:' or 'u', we have to add an '-e' to it in order to combine the next noun or the next adjective (notice that in Persian the adjectives precede the nouns). Consider the noun phrase *a:hu-e nr* (red deer or stag) in which '–e' is the realization of ezafe between the two nouns. In the most other cases the ezafe is pronounced as 'e' but it is not written.

Adjectives follow the same morphological patterns as nouns. They can also appear with comparative and superlative morphemes. Certain adverbs, mainly manner adverbs, can behave like adjectives and can appear with all the adjectival affixes [11].

The inflectional system for the Persian verbs consists of simple forms and compound forms; the latter are forms that require an auxiliary verb. The simple forms are divided into two groups according to the stem they use in their formation, present or past. The citation form for the verb is the infinitive [11].

Although inflectional morphemes can attach to the verbs for conjugation, they also may attach to the end of the nouns or adjectives and convey a complete sentence enclosing subject and verb. However, in these cases the verb can only be intransitive and in the form of "be". Consider the following example:

f*qir* (poor) + id = *fgirid* (you are poor.)

Persian uses the Arabic alphabet. Texts are written from right to left. Vowels generally known as short vowels (a, e, o) are usually not written; only the long vowels (i, u, a:) are represented in the text. The only exception is for short vowels occurring at the initial position of a word, in the case of which the short vowels are written. Consider the following examples in which the initial short vowels are transcripted:
-ota:q (room), omid (hope), ordk (duck);
-ensa:n (human), ektsha:f (discovery), eh'tma:l (probability);

DOI: 01.IJIT.01.01.5

-ana:r (pomegranate), axba:r (news), abza:r (tools).

In Persian writing system letters in a word are often connected to each other. Most characters have two or three alternative forms depending on their position within the word. The initial form indicates that no element is attached to the element from the right. Note that an initial form does not mean that the character is in the beginning of a word, it only indicates that the character is not at the end of the word. Characters are in medial form if they have an attaching character both before and after them. The final form denotes that the character is at the end of a word. The final forms can therefore be used to mark the word boundaries. However, certain characters (see Appendix I) have only one form regardless of their position within the word.

In written text, words are usually separated by a space. Compounds and detachable morphemes (i.e., morphemes following a word ending in final form character), however, are written without a space separating them. In other words, the two parts of a compound appear next to each other but the first element in the compound will usually end in a final form character, hence it would be possible to recognize the two parts of the compound. This form is not very consistent, however, and sometimes words can appear without a space between them. If the first word ends in a character that has a final form, then we can easily distinguish the word boundary. But if the first word ends in one of the characters that have only one form, the end of the word is not clear. Although this latter case is usually avoided in written text, it is not rare. Furthermore, a space is sometimes inserted between a word and the morpheme. In such cases, the morpheme needs to be reattached (or the apace eliminated) before proceeding to morphological analysis of the text [11].

III.CORPUS

The training corpus used for POS disambiguation of Persian texts contains 40,000 words, 2120 sentences. This corpus, which are extracted from Persian Ettela'at

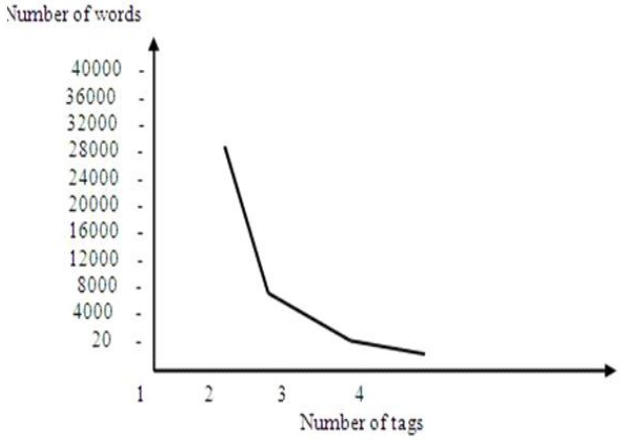

Figure 1. Distribution of words with regard to the number of tags

Newspaper date 12 March 2004 in different fields as economics, politics, culture, sport and science has manually been tagged, i.e. a genotype as well as a genotype decision were considered for each word. These genotypes have from one to five members that correspond to words having from one to five possible tags, respectively. Distribution of words in the corpus according to the member numbers of each genotype has been shown in Figure 1. As it is clear, about 74.85% of the words have only one member in their genotype, and thus unambiguous, about 20% of the words have two members, about 4.6% of the words have three members, and about 0.55% of the words have four and five members in their genotype. Most frequent parts of speech in our corpus have been represented in Table 1. As Table 1 shows, nouns compose one-third proportion of the corpus and prepositions occupy a relatively significant space in Persian text most of which are ambiguous between being a noun and a preposition.

The POS tagset used for annotating Persian corpus distinguishes 62 tags each of which consists of from two to four characters. In linguistic analysis using one or the other tagset depends on its application in the related task. Hence, for working with genotype frequencies in our corpus we constructed a tagset consisting of 62 tags based on Persian morphological characteristics and requirements of the current experiment (see Appendix II). For example, in this tagset we distinguish the two tags [nsg] and [nsgz], Which stand for 'singular common noun' and 'singular common noun plus indefinite suffix 'i',since in Persian there are a relatively large number of words whose genotype contains [nsgz] and genotype decision is another tag such as adjective, simple noun or verb. There are three single-member tags in our tagset as [ra] [xud] and [ke]. Table 2 shows the most frequent genotypes and their respective frequencies in our corpus. The test corpus used for evaluating the experiment results, consists of 6700 Persian words. Each word in the test corpus has already been received a genotype according to a lexicon extracted from Moin Dictionary of Persian words [5] consisting of all Persian words with their different parts of speech.

TABLE I.
MOST FREQUENT PARTS OF SEECH IN THE TRAINING CORPUS

Part-of-speech occurrence	probability
noun: [nsg] [nsgz] [npl] [nplz]	34.55%
preposition: [pr] [prc]	12.65%
adjective: [aj] [ajc] [ajs]	10.4%
verb: [vl] [vx] [vd] [vf] [vsp] [vsu] [vprs] [vim] [vpp] [vss] [vppo] [vfss] [vin]	7.8%
conjunction: [cjc] [cjs]	5.3%

DOI: 01.IJIT.01.01.5

TABLE II.
MOST FREQUENT GENOTYPE IN THE CORPUS

N	genotype	frequency	occurrence probability of each tag
1	[nsg] [pr]	2003	[nsg] 4.74% , [pr] 96.85%
2	[aj] [nsgz]	1396	[aj] 85.38% , [nsgz] 14.61%
3	[aj] [nsg]	1265	[aj] 47.82% , [nsg] 52.17%
4	[nsg] [vsp]	381	[nsg] 37.27% , [vsp] 62.72%
5	[aj] [nsg] [nsgz]	360	[aj] 16.66% , [nsg] 77.5% , [nsgz] 5.83%
6	[aj] [av]	284	[aj] 70.77% , [av] 29.22%
7	[aj] [av] [nsg]	223	[aj] 44.84% , [av] 51.56% , [nsg] 3.5%
8	[aj] [nsg] [vsu]	120	[aj] 12.5% , [nsg] 85% , [vsu] 2.5%
9	[nsg] [nsgz]	102	[nsg] 86.27% , [nsgz] 13.72%
10	[aj] [nsg] [nsgz] [vl]	100	[aj] 19% , [nsg] 70% , [nsgz] 3% , [vl] 8%

II.EXPERIMENT

As it has mentioned in the previous section, we assign a genotype as well as a genotype decision for each word in our training corpus. We perform the experiment with the collected data in two levels. During the first level in which only unigrams of genotypes are considered, we assign a tag for each word of the test corpus (consisting of 6,700 words) which has the highest occurrence probability in the genotype to which the word belongs according to the genotype decision of the training corpus. Consider, for instance, the word *ba:zrga:ni* in the test corpus, which belongs to the genotype {[aj] [nsg] [nsgz]}. Considering our training corpus, in this single genotype the member [nsg] has the highest probability of occurrence as the correct tag for the words belonging to this genotype regardless of what the word is. It means, in about 77.5% of cases [nsg] (singular common noun) is the correct tag, in about 16.66% of cases [aj] (adjective), and in about 5.83% of cases [nsgz] (singular noun plus indefinite suffix). Hence, we assign [nsg] as the correct tag for the word *ba:zrga:ni*. However, using unigram genotypes cannot be stable and dependable, and leads to some incorrect decisions in disambiguation due to its overgeneralizations. If we consider third column of numbers 3 and 7 in table 2, we can infer that the occurrence probability of two tags [aj] and [nsg] are very close to each other so that selecting each of them based on unigram genotypes level cannot be reliable.For solving this kind of problems we developed our experiment in the second level, i.e. decision based on bigram genotypes. In this level we consider the right context of the genotype in determining the correct tag for each word.

It is believed that, using context for disambiguation increases the accuracy of the results. To show the effectiveness of using context in assigning the correct tag for words, we selected a typical ambiguous word, *gsha:ishi*, extracted from the following sentence of the test corpus:

Pzhuhshgra:n tva:nsth-and ka:rkrd ba:zrga:ni dr gz"'shth ra: dgrgun knnd v dr a:n gsha:ishi pdid a:vrnd.

'Researchers could change the trade output in the past and make an improvement on it.'

Which belongs to the genotype {[aj] [nsgz]}. If we consider the unigram genotypes for determining the correct tag for this word, naturally we must assign [aj] (adjective) as the correct tag for it based on its highest occurrence probability over [nsgz] in our training corpus. Now we consider the immediate context to which the underlined word belongs from the right (here it occurs in the left side because we transcript the Persian sentence with English letters.) which is *a:n* (it). Calculating the frequency of occurring genotype {[ajd][pd]} with genotype decision [pd] before genotype {[aj] [nsgz]}, the correct tag for the word *gsha:ishi* turns to be [nsgz] which is the correct tag for this word in this sentence. In the same sentence, there are three other ambiguous words (*ka:rkrd, ba:zrga:ni, gz"'shth*), disambiguation process of which using immediate right context results in desirable tag decision, while using unigram genotypes leads to incorrect tag decision as it is shown in Table 3:

TABLE III.
UNIGRAM GENOTYPE DISAMBIGUATION LEVEL VS. BI-GRAM ONE

word	genotype	unigram	bigram	genotype decision
ba:zrga:n	[aj] [nsg] [nsgz]	[nsg]	[aj]	[aj]
gz``shth	[aj] [av] [nsg] [nsgz]	[aj]	[nsg]	[nsg]
ka:rkrd	[nsg] [vsp]	[vsp]	[nsg]	[nsg]
gsha:ishi	[aj] [nsgz]	[aj]	[nsgz]	[nsgz]

DOI: 01.IJIT.01.01.5

As Table 3 shows there are three ambiguous words in one sentence for which using context for disambiguation gives better results than using the highest probable tag in the limits of one and the same genotype in the corpus. It should also be stated that in this sentence there are some other ambiguous words as: *ba:zrga:ni, dr, dr* and *a:n* for which there is no difference between decisions of unigram and bigram genotype disambiguation.

Tzoukermann and Radev suggest a third level of disambiguation as trigram genotypes disambiguation which considers not only the right context of the ambiguous word but also the left context to assign the correct tag for that word. They claim that using trigram genotypes leads to more accurate results in POS disambiguation of the word [18]. The fact is that, although it is obvious that using trigram genotypes can assign more accurate tag for the given word, finding enough tagging sequences for calculating frequencies and then most probable case for such a complicated context calls for a larger training corpus at least for a language like Persian with sparse data.

I believe that to improve theaccuracy of our method for POSdisambiguation of Persian text using trigram level we require at least a 200,000 words corpus annotated manually for genotype and genotype decision for each word.

V.RESULTS

We performed the tagger generation process on a 40,000 words training corpus of Persian extracted from the Ettela'at Newspaper date 12 March 2004, and tested on 6,700 test words extracted from the same Newspaper but in different dates in two stages. In the first stage we used unigram genotypes disambiguation and in the second stage we used bigram genotypes disambiguation. In both stages we were concerned with statistical estimations on the genotypes only, regardless of the words. Hence, the absence of an individual test word from the training corpus does not affect our estimations and thus the results. Table 4 demonstrates the accuracy rate of the two levels of disambiguation in comparison.

TABLE IV
COMPARING ACCURACY OF THE TWO LEVELS OF DISAMBIGUATION

	total number of words	ambiguous words	correctly tagged words	accuracy
Unigram level	6700	23560	13066	55.2%
Bigram level	6700	23560	20675	87.4%

As Table 4 shows, the accuracy rate of disambiguation based on genotypes of individual words is considerably lower than the accuracy rate gained from the second stage of experiment, i.e. using the immediate right context of genotype to the given word. In fact, unigram genotypes disambiguation is a shallow estimation on which one cannot rely as an efficient disambiguation procedure, though working with it is much easier than the alternative choice.

VI.COMPARISON WITH OTHER METHODS

In most statistically-based methods for POS disambiguation including those mentioned in this paper, the basis of counts or occurrence frequencies is on the words themselves and not on their tags or genotype. It is natural that, calculating a certain number of genotypes and genotype decisions is much easier than relying on the number of certain words and working with individual words, since the former is very smaller in number size. Moreover, using genotypes than words for calculating frequencies has the advantage of calling for a small training corpus needed for manual annotation of the words. This property is valuable and timely in languages including Persian for which a large available tagged corpus does not exist.

The method presented in this paper is somehow similar to the method proposed by [20] who also used word class for POS disambiguation in main concepts. However, our method is different from theirs in some aspects. First, they referred and used the term 'decision genotype' for the tag belonging to a word genotype with the highest probability. In our experiment we found it more applicable to use decision genotype as the correct tag for the given word in a given context. Taking decision genotype as the correct word tag rather than the most probable tag can help us to better apply this model in the experiment, since, as we know the most probable tag of a word is not necessarily the correct tag for that word. Second, dealing with bigram genotypes disambiguation level, we calculate co-occurrence probability of the given word genotype and decision genotype of its immediate right word, while Tzoukermann and Radev counts co-occurrence probability of the genotypes of the two words. In case that we count decision genotype of the right word of the given word, we will deal with much more instances to be calculated for frequency and as a result, a smaller corpus is needed.

II.CONCLUSION

In this experiment we used a rather different method of statistical POS disambiguation in the Persian language which has already been applied in French with some minor differences. The results gained from the experiment indicate that the method is also applicable in Persian and the accuracy rate of the second level of disambiguation is promising with regard to the fact that there is no large annotated corpus for Persian.

DOI: 01.IJIT.01.01.5

One of the advantages of this method among others is that in our system all statistical estimations are entirely done on the genotypes only, regardless of words, hence, handling new words that have not been seen in the training corpus will be possible without any difficulty. This property has also the advantage of ability to smooth probabilities. Moreover, estimating probabilities on a small number of genotypes rather than a large number of words is an enormous gain.

As it has been mentioned earlier, using a third level of disambiguation considering right as well as left context to which the ambiguous word belongs can increase the accuracy rate of our experiment to a high degree. However, finding three immediate genotype sequences in such a small corpus like ours with sufficient frequencies to work with is a rather impossible task. We estimate that performing with this third level of disambiguation calls for preparing an annotated corpus of Persian of the size 200,000 words. This can be a further direction for developing this disambiguation method for Persian.

APPENDIX I

Persian transcription table:

English Transcription	Key word	Persian example
a:	Calm	*a:hu* (beer)
a	Bad	*aql* (reason)
e	Bed	*eshq* (love)
o	For	*olgu* (pattern)
b	Back	*abr* (cloud)
p	Pen	*pa:iin* (below)
t	Tea	*tnha:* (alone)
s`	Soon	*s`ria* (heaven)
j	Jump	*jha:n* (world)
ch	Cheer	*chha:r* (four)
h`	Hot	*eh`sa:n* (benevolence)
x	_____	*xa:nh* (home)
d	Day	*sd* (dam)
z```	Zero	*a:z```in* (decoration)
r	Red	*sfir* (ambassador)
z	Zero	*zur* (force)
zh	Pleasure	*mzhdh* (presage)
s	Soon	*rsid* (receipt)
z`	Zero	*nhz`t* (movement)
t`	Tea	*t`ahr* (clean)
z``	Zero	*z``hur* (advent)
?	_____	*?lm* (science)
gh	_____	*gha:ib* (absent)
f	Fat	*Eftxa:r* (honour)
q	_____	*eshq* (love)
k	Key	*mh`km* (firm)
g	Get	*grft* (got)
l	Lead	*mbl* (furniture)
m	Sum	*mn* (I)
n	Sun	*zmin* (earth)
u, v	very, boot	*avl, xub* (first, good)
h	Hot	*bhtr* (better)
i	Happy	*binsh* (insight)

DOI: 01.IJIT.01.01.5

APPENDIX II

Tagset used in the experiment:

N	Tag	Description	Example
1	[aj]	General adjective	*ziba:*
2	[ajc]	Comparative adjective	*bhtr*
3	[ajd]	Demonstrative adjective	*in, a:n*
4	[ajp]	Adjective plus objective pronoun	*xnh-ash*
5	[ajs]	Superlative adjective	*bhtrin*
6	[ajz]	Adjective plus indefinite suffix	*qshng-i*
7	[ate]	Attribute exclamation	*?jb*
8	[ats]	Attribute subordinate	*inkh*
9	[av]	General adverb	*sri?*
10	[avp]	Adverb of place	*Inja:*
11	[avq]	Adverb of interrogative	*chgunh*
12	[avt]	Adverb of time	*diruz*
13	[aya]	Question word equivalent to 'do', 'if'	*a:ya*
14	[cjc]	Coordinate conjunction	*v, ama:*
15	[cjs]	Subordinate conjunction	*agrchh*
16	[co]	Colon	*:*
17	[cr]	Cardinal number	*pnj*
18	[cru]	Cardinal number unspecified	*dhha:*
19	[fs]	Full stop	*.*
20	[dti]	Indefinite determiner	*hr*
21	[dtq]	Interrogative determiner	*kda:m*
22	[itj]	Interjection or other isolate	*blh, ha:*
23	[ke]	Relative pronoun	*ke*
24	[ltr]	Letter	*b*
25	[np]	Noun plus objective pronoun	*kta:b-m*
26	[npl]	Plural common noun	*mizha:*
27	[nplz]	Plural common noun plus indefinite	*drxta:n-i*
28	[npr]	suffix	*Ali*
29	[nsg]	Proper noun	*qa:li*
30	[nsgz]	Singular common noun	*drxt-i*
31	[or]	Singular common noun plus indefinite	*hftm*
32	[pd]	suffix	*in, a:n*
33	[pni]	Ordinal numeral	*ksi*
34	[pnp]	Demonstrative pronoun	*shma:*
35	[pnr]	Indefinite pronoun	*xodsh*
36	[pnrc]	Personal pronoun	*hmdigr*
37	[pq]	Reflective pronoun	*chh, kh*
38	[pr]	Reciprocal pronoun	*az, ba:*
39	[prc]	Interrogative pronoun	*birun az*
40	[prt]	Preposition	*rfth*
41	[pul]	Compound preposition	*([*
42	[pun]	Past participle	*, ?*
43	[pur]	Punctuation, left bracket	*)]*
44	[pv]	Punctuation	*tust*
45	[ra]	Punctuation, right bracket	*ra:*
46	[unc]	Pronoun plus linking verb	*internet*
47	[vd]	Direct object marker	*mirvd*
48	[vf]	Unclassified item (non-Persian word)	*xa:hm rft*
49	[vfss]	Verb, declarative	*gfth xa:hd*
50	[vim]	Verb, future	*shd*
51	[vin]	Verb, passive future	*bnshin*
52	[vl]	Verb, imperative	*shnidn*
53	[vp]	Verb, infinitive	*ast, bud*
54	[vpp]	Verb, linking	*gftm-sh*
55	[vppo]	Verb plus objective pronoun	*gfth budm*
56	[vprs]	Verb, past perfect	*gfth ba:shm*
57	[vsp]	Verb, past potential	*nshsth am*
58	[vss]	Verb, present perfect	*nvshtnd*
59	[vsu]	Verb, simple past	*shnidh shd*
60	[vx]	Verb, passive	*brvd*
61	[xud]	Verb, subjunctive	*mitva:nd*
62	[??]	Verb, auxiliary or modal	*xud*
		Emphatic form without specifying person	_____
		Unknown items	

DOI: 01.IJIT.01.01.5

It should be stated that, the two letters *v* and *u* have the same surface form in Persian transcription but different pronunciations depending on the word in which they occur. *u* is a long vowel, while *v* is a consonant. Hence, in this paper we used these two letters differently in transcription to show their different pronunciations.

REFERENCES

[1] S. M. Assi, and M. Haji Abdolhosseini, "Grammatical tagging of a Persian Corpus". *International Journal of Corpus Linguistics*, 5 (1), 69-81, 2000.

[2] M.Braschler, and P. Schauble, "Using corpus-based approaches in a system for multilingual information retrieval". *Information Retrieval*, 3, PP. 273-284, 2000.

[3] E. Brill, "Unsupervised learning of disambiguation rules for part of speech tagging". In *2nd Workshop on Large Corpora*, Boston, USA, 1995.

[4] E. Brill, "A simple rule-based part of speech tagger". In *proceeding of 3rd Conference on Applied Computational Linguistics*, Trento, Italy, PP. 152-155, 1992.

[5] S. M. Conrad, "The importance of corpus-based research for language teachers". *System*, 27, PP. 1-18, 1999.

[6] D. Cutting,; J. Kupiec,; J. Peterson, and P. Sibun. "A practical part of speech tagger". In *proceeding of 3rd Conference on Applied Computational Linguistics*, Trento, Italy, PP. 133-140, 1992.

[7] W. Daelemans,; J. Zavrel,; P. Berck, and S. Gilis, "MBT: A memory-based part of speech tagger-generator". In *4th Workshop on Very Large Corpora, Special Interest Group for Linguistic Data and Corpus-based Approaches (SIGDAT) of the ACL*, Copenhagen, Denmark. PP. 14-27, 1996.

[8] H. Givi, and H. Anvari, "*Persian grammar, (1) and (2)*". Fatemi Press, Tehran, 1998.

[9] A. Kempe, "Part of speech tagging with two sequential transducers". In *proceeding of CLIN 2000*, Tilburg, the Netherlands, November 3, PP. 88-96, 2000.

[10] A. Masayuki. "Corpus-based Japanese morphological analysis". *Doctor's Thesis*, Nara Institute of Science and Technology, NAIST-IS-DT0161001, 2003.

[11] K. Megerdoomian, "Persian computational morphology: A unification-based approach". *Computing research Laboratory*, New Mexico State University, New Mexico, 2000.

[12] M. Moin, "*Persian Dictionary of Moin*". Moin Press, Tehran, 2002.

[13] T. Mosavi Miangah, "Automatic lemmatization of Persian words". *Journal of Quantitative Linguistics*, Vol. 13, No. 1, pp. 1-15, 2006**.**

[14] T. Mosavi Miangah, and A. Delavar Khalafi, "Word sense disambiguation using target language corpus in a machine translation system". *Literary and Linguistic Computing*, 20(2). PP. 237-249, 2005.

[15] T. Mosavi Miangah, "Problems of English-Persian machine translation". *Journal of Philology*, 3(12), PP. 38-42, 2002.

[16] T. Mosavi Miangah, "Comparative analysis of noun phrase for MT (with reference to English and Persian)." *Problems of Language Theory and Translation Science*, Moscow Pedagogical University, 6, PP. 68-78, 2001.

[17] G. Petasis,; G. Paliouras,; V. Karkaletsis, and C. D. "Spyropoulos, Resolving part of speech ambiguity in the Greek language using learning techniques". In *proceeding of the ECCAL Advanced Course on Artificial Intelligence (ACAT'99)*, Chania, Greece, 1999.

[18] H. Schuetze, "Distributional part-of-speech tagging. From texts to tags: Issues in Multilingual Language Analysis". Online *proceedings of the ACL SIDGAT Workshop*. On the Internet at http:/ / xxx.lanl.gov/find/cmp-lg, 1995.

[19] J. Tsutsumi, et al., "Multilingual system of machine translation based on statistical information". (in Russian). In *proceeding of QUALICO-9*, 1994.

[20] E. Tzoukermann, and D. R. Radev, "Using word class for part of speech disambiguation". In *4th Workshop on Very Large Corpora, Special Interest Group for Linguistic Data and Corpus-based Approaches (SIGDAT) of the ACL*, Copenhagen, Denmark. PP. 1-13, 1996.

DOI: 01.IJIT.01.01.5